AFOSR-TR- 81-0134

The University of Connecticut
SCHOOL OF ENGINEERING
Storrs, Connecticut 06268

LEVEL

AD A096004

Department of

Electrical Engineering and Computer Science

FILE COPY

81 2 27 065

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
AFOSR-TR-81-0134	AD-A096004	

4. TITLE (and Subtitle)	5. TYPE OF REPORT & PERIOD COVERED
GENERAL CONSTRUCTIVE THEORY OF PARAMETRIC AND ROBUST DATA SMOOTHERS	Interim rept.
	6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)	8. CONTRACT OR GRANT NUMBER(s)
P. Papantoni-Kazakos	AFOSR-78-3695

9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
University of Connecticut Department of Engineering Storrs, CT. 06268	61102F 2304/A5

11. CONTROLLING OFFICE NAME AND ADDRESS	12. REPORT DATE
Air Force Office of Scientific Research/NM Bolling AFB, Washington, DC 20332	December 1980
	13. NUMBER OF PAGES
	31

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report)
	UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

EECS-TR-80-11

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

In relatively recent years, several algorithms for smoothering of time series have been proposed by statisticians. Some of the simpler such algorithms have been also applied in several engineering applications such as Image Processing. The smoothering problem and the implied objective have not been formalized and stated, however. This fact presents a serious handicap when different smoothing algorithms are to be compared in terms of their performance. In this paper we take a fresh and daring approach to the whole smoothing problem. We formalize

the problem as the extraction of a low entropy process form a high
entropy process, and as a result we present a constructive theory
of parametric and robust data smoothers. We claim that parametric data
smoothers are analog-to-digital converters, and that robust data smoothers
are stochastic such converters.

GENERAL CONSTRUCTIVE THEORY OF PARAMETRIC

AND ROBUST DATA SMOOTHERS

P. Papantoni-Kazakos

Technical Report EECS-TR-80-11

December, 1980

General Constructive Theory of Parametric

and Robust Data Smoothers[*]

P. Papantoni-Kazakos
The University of Connecticut
Dept. of Electrical Engineering
and Computer Science
Storrs, Connecticut   06268

Abstract

> In relatively recent years, several algorithms for smoothing of time series

have been proposed by statisticians. Some of the simpler such algorithms have

been also applied in several engineering applications such as Image Processing.

The smoothing problem and the implied objective have not been formalized

and stated, however. This fact presents a serious handicap when different

smoothing algorithms are to be compared in terms of their performance.

In this paper we take a fresh and daring approach to the whole smoothing

problem. We formalize the problem as the extraction of a low entropy process

from a high entropy process, and as a result we present a constructive theory

of parametric and robust data smoothers.

We claim that parametric data smoothers are analog-to-digital converters,

and that robust data smoothers are stochastic such converters.

1. Introduction

Whenever a new experiment is set, there is no a priori information as to the statistical nature of the outcome. It is desirable, therefore, to attempt conclusive evaluation of the statistical behavior of the experiment outcomes, through repetitive observations from the same experimental setting. The task is particularly challenging due to the noisy nature of the observations.

In satellite communications, the transmitted useful data are corrupted by atmospheric noise. The task of the receiver is to extract the unknown useful data from their noisy version. The encoded transmission of images through satellite, fall into this category.

The above, are only two of many indicative applications, which initiated the consideration of data smoothers. The general objective of the data smoothers is the extraction of an unknown data process (or time-series as called in the statistical literature) from its noisy version.

Unfortunately, no qualitative definition of the data smoothers exists. The objective of the data smoothers has been only implicitly determined in terms as general as: "Extraction of a smooth data process from a nonsmooth noisy data sequence".

Using the above vague definition of the objective of the data smoothers, Tukey [1] proposed several ad hoc smoothing algorithms. Due to the lack of a qualitative theory, Tukey's algorithms can not be evaluated coherently. Recognizing this problem, Mallows [3,4] presented a first formalization of some theory of nonlinear smoothers. Mallow's approach consists of guaranteeing some desirable properties, it is limited to memoryless data, and it still lacks general qualitative formalization. Martin [5] assumed certain known structure of the data

process and presented accordingly a class of conditional – mean type robust data smoothers. On the other hand, recognizing the need for data smoothers on unknown data processes, Huber [2] took a statistics-free approach and formalized robust smoothers within the framework of data splines.

Our view is that the approaches in [2-5] are enlightening but also intriguing. A qualitative formalization of the smoothing problem is still lacking, and this fact makes the evaluation of different proposed data smoothers virtually impossible. Furthermore, we are in agreement with Huber's general philosophy. Specifically, we think that in the smoothing problem it is not realistic to assume certain given structure of the data process. The very objective of the data smoothers is the extraction of the data sequence and the possibly consequent evaluation of its underline statistical behavior.

We believe that the key to a general qualitative theory of data smoothers is the qualitative characterization of the term "smooth". Indeed, the term implies certain general knowledge about the process which generates the data sequence to be extracted. In fact, we believe that the term implies possibly relative knowledge about the data process, in comparison to the noise process by which the data process is corrupted.

From now on, we will use the following terminology: We will call informa- tion carrying process, the stochastic process which generates the data sequence we wish to extract. We will call noise process, the stochastic process which corrupts the data from the information carrying process. We call observation process the stochastic process induced by the conjuction of the information carrying and the noise processes.

Starting the introduction of our theory, it is necessary to refer to the implicit assumption under the operation of the data smoothers. The assumption

is that the information carrying process allows for data sequences with only

low level variations in time, and that any high level such variations are due

to the noise process. This assumption is hidden behind the term "smooth"

assigned to the information carrying process. But in more qualitative terms,

this implicit assumption equivalently means that the uncertainty about the

information carrying process is low, as compared to the uncertainty about the

noise process. Using the term entropy for uncertainty and the fact that the

process induced by a low entropy process in conjuction with a high entropy pro-

cess is a high entropy process (the observation process in this case), we can

formalize the objective of the data smoothers in the following way:

> Extract some data sequence generated by a low entropy process, from
>
> a data sequence generated by a high entropy process.

The above formalization regarding the objective of the data smoothers is

still vague, but it sheds light as to the proper direction towards a qualitative

general theory. Taking one further step towards the development of such a theory,

we first consider parametric data smoothers and then we expand to incorporate

consideration of robust data smoothers.

We consider the problem of parametric data smoothers arising when the noise

process is well-defined, or more specifically, when every process induced by

any given data sequence from the information carrying process and the noise

process, is well-defined.

We consider the problem of robust data smoothers arising, when every pro-

cess induced by any given data sequence from the information carrying process

and the noise process, is statistically contaminated.

In both the problems of parametric data smoothers and robust data smoothers,

we assume that the information carrying process is a low entropy process as

compared to the noise process. To quantify this, we may assume an upper bound

on this entropy, in which case we may proceed with the following definitions:

Definition 1

A stochastic process is <u>smooth with degree of smoothness β</u> if its entropy

is bounded from above by β.

Definition 2

A stochastic process is <u>oversmooth</u> if its entropy is zero.

An oversmooth process corresponds to a constant, thus definition 2 is

consistent with Huber's oversmoothing [2].

Also, the assumption of smoothness of some degree is implicit behind all

the proposed smoothing operations [1].

In this paper, we will consider stationary and ergodic information carrying

and noise processes. We will not restrict our analysis to memoryless processes,

however.

2.   Preliminaries

Let $[\mu,A,X]$ be some stationary and ergodic process, where μ is the proba-

bility measure of the process, A is the process alphabet, and X its name. Denote

by x a discrete-time, infinite sequence of process elements, and by x^n an

n-dimensional sector of x. Denote also by X_i^j; $j \geq i$ the sequence X_i, X_{i+1}, ..., X_j

of j-i+1 consequtive elements from the process.

Let $[\nu,B,Y]$ be another stationary and ergodic process, where ν,B,Y,y,y^n,

Y_i^j; $j \geq i$ as above.

Let ν_x be the measure induced by an infinite sequence x from the process

$[\mu,A,X]$ and the process $[\nu,B,Y]$, and let C, W be the induced alphabet and name

respectively. We will assume that the superposition relationship between the [μ,A,X] and [ν,B,Y] processes is well-known, and that therefore if x known and [ν,B,Y] well-known then the process [ν$_x$,C,W] is well-known. The process [ν$_x$,C,W] can be looked upon as a transmission channel for the process [μ,A,X], and the superposition relationship between the processes [μ,A,X] and [ν,B,Y] can be, for example, additive.

Let us denote by w, w^n, W_i^j ; $j \geq i$ sequences from the process [ν$_x$,C,W], as for the process [μ,A,X].

Let us denote by $\{s_\ell\}$ a sequence of sliding block functions on data sequences from the process [ν$_x$,C,W], where the sequence is determined by varying the length ℓ of the sliding block window.

Each sliding block function $s_\ell(\cdot)$ produces a discrete-time sequence z with elements Z_j ; $j = \ldots, -1, 0, 1,\ldots$, through the following operation:

$$Z_j = s_\ell \left(W_{j-\ell+m+1}^{j+m} \right) \; ; \; m+1 \leq \ell \tag{1}$$

The function $s_\ell(\cdot)$ in (1) is time-invariant, it operates on ℓ-length blocks of elements from the process [ν$_x$,C,W], and it can be either a deterministic or a stochastic function. If $\nu_x \, s_\ell^{-1}$ is the measure induced by ν_x and $s_\ell(\cdot)$ for different w sequences, if D is the induced alphabet, and if [ν$_x$,C,W] is stationary and ergodic, then so is the process [$\nu_x \, s_\ell^{-1}$, D,Z] [see reference 7]. We will denote by s_{ℓ,w^ℓ} the measure induced by the sequence w^ℓ and the sliding block function in (1). This last measure is nontrivial if the sliding block function $s_\ell(\cdot)$ is stochastic.

Given the two stationary and ergodic processes [μ,A,X] and [ν,B,Y], we will consider superposition relationships between them such that they result

in a process $[\mu\nu, E, V]$ which is also stationary and ergodic. Such a superposition relationship is the additive. Then, the class of processes $[\nu_x, C, W]$ generated by different x sequences from the process $[\mu, A, X]$, are all stationary and ergodic also. In this case, conclusions about the measure ν_x can be drawn from time-domain averages on the sequence w. Also, if the process $[\nu, B, Y]$ is an infinite alphabet (analog), high entropy process, and the process $[\mu, A, X]$ is a low entropy process, the process $[\mu\nu, E, V]$ and the class of processes $[\nu_x, C, W]$ will all be infinite alphabet and high entropy processes.

If we wish to reduce the entropy of the process $[\nu_x, C, W]$ to values below a finite value β, quantization is necessary.

In the following sections, we will use the construction of empirical measures from data sequences, as in [6,7]. Specifically, given a finite discrete-time sequence w^n from the stationary and ergodic process $[\nu_x, C, W]$, we form a string $w = (\ldots, w^n, w^n, ---)$ by repeating the sequence w^n. If μ_{w^n} is the empirical measure formed by assigning probability n^{-1} on each string $T^i w$; $i = 0, 1, \ldots, n-1$, where T indicates one step shift in time, then:

$$\mu_{w^n}(F) = \sum_{i:T^i w \varepsilon F} n^{-1} \quad ; \ F \ \varepsilon \ A^\infty \qquad (2)$$

; where A^∞ the generalization of the σ-algebra A of sets on the space C on which each datum W_i assumes values. The empirical measure μ_{w^n} has restrictions $\mu^k_{w^n}$ which assign measure n^{-1} to each k-tuple of k adjacent symbols within w^n. The process in (2) is then periodic, stationary, ergodic, and the empirical measures $\mu^k_{w^n}$ are trustworthy for $k \leq n$. Furthermore, if Π_k indicates the k-dimensional Prohorov distance as in [6,7], then the ergodic theorem implies that for fixed k:

$$\lim_{n\to\infty} \Pi_k\left(\mu^k_{w^n}, \nu^k_x\right) = 0 \ , \ \text{a.e. in measure } \nu_x \qquad (3)$$

; where ν^k_x the k-dimensional restriction of the measure ν_x.

3. Parametric Data Smoothers

a. Model Introduction

Let $[\mu,A,X]$ be the stationary and ergodic information carrying process. This process is unknown, the only knowledge available being that it is a smooth process with degree of smoothness β (as in definition 1).

Let $[\nu,B,Y]$ be the stationary and ergodic noise process. This process is well-known, it is in general an infinite alphabet (analog) process, and its entropy is much larger than the bound β.

To avoid unnecessary complications in our analysis, we will assume [as in 1-5] that the superposition relationship between the processes $[\mu,A,X]$ and $[\nu,B,Y]$ is additive. Then, the observation process $[\mu\nu,E,V]$ induced by the above information carrying and noise processes, is also stationary and ergodic. Furthermore, the process $[\mu\nu,E,V]$ is an infinite alphabet (analog), high entropy process.

We will state the objective of the parametric smoothers formally, through a proposition.

Proposition 1

The objective of the parametric smoothers, is to extract the smooth with degree of smoothness β but otherwise unknown process $[\mu,A,X]$ from sequences v from the observation process $[\mu\nu,E,V]$, where the noise process $[\nu,B,Y]$ is a well-known, high entropy process.

To accomplish the objective of proposition 1, we must perform certain operations on the observation sequence v. Such operations will be called smoothing functions. Also, the observation sequence is usually finite, of say length n. Then, the smoothing functions should operate on finite observation sequences v^n.

The imposed finite sequence limitation, necessitates the most effective utilization of the observation sequence v^n. Due to the stationarity and ergodicity of the observation process $[\mu v, E, V]$, such most effective utilization of the observed data can be obtained through the introduction of a sequence $\{s_\ell\}$ of sliding block smoothing functions. In fact, if $[v_x, C, W]$ is the process induced by some sequence x from the information carrying process and the noise process $[v, B, Y]$, the observed sequences w^n are utilized effectively by a sequence of smoothing functions as in (1), for $\ell \leq n$.

We formalize the above discussion by a proposition.

<u>Proposition 2</u>

The objective of proposition 1 is most effectively accomplished through, in general, a sequence $\{s_\ell\}$; $\ell \leq n$ of sliding block smoothing functions on observed sequences w^n.

It is well known [7,8] that each sliding block smoothing function s_ℓ, when operating on sequences w^n from the stationary and ergodic process $[v_x, C, W]$, induces a stationary and ergodic process. Therefore, time averages converge then asymptotically to probability measures (in measure).

Each of the processes $[v_x \, s_\ell^{-1}, D, Z]$, induced by the process $[v_x, C, W]$ and the sliding block smoothing function s_ℓ, is designed to "approximate" the information carrying process $[\mu, A, X]$. Therefore, each of the processes $[v_x \, s_\ell^{-1}, D, Z]$ should have entropy close to the entropy of the process $[\mu, A, X]$. But if the process $[\mu, A, X]$ is smooth with degree of smoothness β, its entropy is bounded from above by the finite number β. Thus, the entropy of $[v_x \, s_\ell^{-1}, D, Z]$ should be within these limits. On the other hand, the entropy of the process $[v_x, C, W]$ is high compared to β. It is then well known [7] that to achieve the

desirable entropy reduction we must include quantization (analog-to-digital conversion) within the operation of the sliding block smoothing function s_ℓ. Therefore,

Lemma 1

The objective of proposition 1 is most effectively accomplished through, in general, a sequence $\{s_\ell\}$; $\ell \leq n$ of sliding block smoothing functions on observed sequences w^n, where each such function s_ℓ is an analog-to-digital converter.

We would like to point out here that the analog-to-digital or quantization operation is equivalent to the operation of a low pass filter, which eliminates bandwidth resulting from the noise only.

Lemma 1 implies that the class of the appropriate parametric sliding block smoothing functions is a class of non-linear functions, which convert analog data to digital data.

More specifically, we propose a sequence $\{s_\ell\}$ such that each s_ℓ operates on the premise of estimating ℓ-dimensional probability masses from the process $[\nu_x, C, W]$. We propose that then these estimates be used for the extraction of the unknown sequence x.

To illustrate quantitatively our proposal let us first assume that the process $[\mu, A, X]$ is oversmooth (definition 2). Then, the information carrying process reduces to an unknown constant C. This constant is also a location parameter if the additive superposition relationship between the processes $[\mu, A, X]$ and $[\nu, B, Y]$ is true. In this case, we propose a sequence $\{s_\ell\}$ of sliding block smoothing functions, which is reduced to one element s_1. Specifically let $s_1(\cdot)$ operate on the observation sequence w^n in the following way:

$$s_1(w_j) = \begin{cases} 0 & ; \ w_j > 0 \\ 1 & ; \ w_j \leq 0 \end{cases} \tag{4}$$

Then, due to the stationarity and ergodicity of the process $[\nu_x, B, Y]$, the average $n^{-1} \sum_{j=1}^{n} s_1(w_j)$ estimates the probability mass of negative values for the one-dimensional restriction ν_x^1 from the measure ν_x. This estimate converges asymptotically to the true probability mass, in measure ν_x^1. If the noise process $[\nu, B, Y]$ is zero mean, and if F_{ν^1} indicates the cumulative distribution implied by its one-dimensional restriction ν^1, then the constant C is estimated as follows:

$$\hat{C}(w^n) = - F_{\nu^1}^{-1} \left(n^{-1} \sum_{j=1}^{n} s_1(w_j) \right) \tag{5}$$

Let us observe that due to the invertible nature of the function $F_{\nu^1}(x)$, the entropy of the variable $\hat{C}(w^n)$ is equal to the entropy of the variable $n^{-1} \sum_{j=1}^{n} s_1(w_j)$, which converges for $n \to \infty$ to zero.

When the information carrying process $[\mu, A, X]$ is not oversmooth but simply smooth with some possibly unknown degree of smoothness, we propose a sequence $\{s_\ell\}$; $\ell \leq n$ of sliding block smoothing functions. We propose, more specifically, that the $s_1(\cdot)$ such function operate as in (4) and (5) for the estimation of the digit x_1 from the sequence x^n. Furthermore, we propose that the $s_2(\cdot)$ smoothing function be used to estimate the two-dimensional probability masses of ν_x^2 conditioned on the estimated by $s_1(\cdot)$ digit x_1, and subsequently be used to estimate x_2 conditioned on the x_1 estimated value. In general, we propose a recursive use of the functions s_ℓ, in such a way that $s_\ell(\cdot)$ is used to estimate the conditional measure $\nu_x^\ell \left(x_\ell \big| x_1^{\ell-1} \right)$, and subsequently the digit x_ℓ conditioned on the previously estimated vector $x_1^{\ell-1}$.

This subsection was dedicated to some explanatory statements regarding the general approach we adopt in this paper. Our approach is further formalized in the following subsection.

b. Rigorous Abstract Formalization

Based on the preceding discussions, we consider ourselves ready at this point, to formalize a general theory on parametric data smoothers.

Let $[\nu, B, Y]$ be a well-known infinite alphabet, stationary and ergodic, high entropy stochastic noise process.

Let $[\mu\nu, E, V]$ be an infinite alphabet, stationary and ergodic, high entropy observation process, induced by a well-known superposition relationship between the well known process $[\nu, B, Y]$ and some unknown low entropy stationary and ergodic information carrying process $[\mu, A, X]$.

Let $\{s_\ell\}$ be a sequence of sliding block functions, operating on sequences v from the $[\mu\nu, E, V]$ process. Each sliding block function $s_\ell(\cdot)$ produces a discrete-time sequence z with elements Z_j through the following operation:

$$Z_j = s_\ell \left(v_{j-\ell+m+1}^{j+m} \right) ; \quad m+1 \leq \ell \tag{6}$$

As in section 2, the function in (6) is time invariant, it operates on ℓ-length blocks of elements from the sequence v, and it can be of either deterministic or stochastic nature. If it is of stochastic nature, s_{ℓ,v^ℓ} will indicate the measure induced by s_ℓ and a given sector v^ℓ from the sequence v. We will denote by $\mu\nu \, s_\ell^{-1}$ the measure induced by the process $[\mu\nu, E, V]$ and the function s_ℓ, for different v^ℓ sequences. We will denote by D the induced alphabet. Then, the process $[\mu\nu s_\ell^{-1}, D, Z]$ is also stationary and ergodic.

We now proceed with the following definition:

Definition 3

The sequence $\{s_\ell\}$ of sliding block functions is a sequence of __parametric data smoothers__ for the process $[\mu\nu,E,V]$, if each induced process $[\mu\nu s_\ell^{-1},D,Z]$ is an estimate of the information carrying process $[\mu,A,X]$.

Let it now be known that the unknown information carrying process $[\mu,A,X]$ is smooth with degree of smoothness β (definition 1). Then, the "estimate-processes" $[\mu\nu s_\ell^{-1},D,Z]$ should be also smooth with degree of smoothers β.

We can now proceed with the following theorem:

Theorem 1

Given that the information carrying process $[\mu,A,X]$ is __smooth with degree of__ smoothness β, and that the observation process $[\mu\nu,E,V]$ is a high entropy process with entropy much higher than β, the sequence $\{s_\ell\}$ of parametric data smoothers for the process $[\mu\nu,E,V]$ is a sequence of sliding block analog-to-digital converters (or quantizers).

The proof of the theorem is straight-forward from the definition of parametric data smoothness for the process $[\mu\nu,E,V]$ (definition 3), and the fact that entropy reduction (from the entropy of the observation process to the entropy of the information carrying process) can be obtained only through quantization.

We will conclude this section by pointing out emphatically that definition 3 and theorem 1 indicate clearly the appropriate performance criterion in the design and evaluation of parametric data smoothers. In particular, this criterion is entropy matching between the processes induced by the smoothers and the entropy of the unknown information carrying process. The choice of this criterion is

clearly dictated by the very nature of the problem. Furthermore, the entropy performance criterion also dictates the result in theorem 1.

4. Robust Data Smoothers

a. Definition

Our analysis in this section is based of course on the theoretical formalization for parametric data smoothers in section 3, and it is parallel to the analysis in [7,8].

We will first explain the statistical model applicable to robust data smoothers.

As with the parametric data smoothing problem, the information carrying process $[\mu,A,X]$ is unknown. What is known, however, about this process, is that it is ergodic and stationary and that it is smooth with degree of smoothness β.

The basic difference between the model of section 3 and the present model, lies on the knowledge about the noise process $[\nu,B,Y]$. Indeed, here we assume that the noise process may be statistically contaminated. That is, it may be any member from a family M of stationary and ergodic processes. As in [7] we will assume that the exact membership of the family M is not known, and we will focus on families of data smoothers which guarantee local performance stability around a nominal well known noise process $[\nu_o,B,Y]$.

As emphasized in section 3, the performance criterion for the design and evaluation of parametric data smoothers should be the output entropy induced. Therefore, when the noise process is ill-defined, the local performance stability sought should be stability in the output entropy induced, when statistical deviations from the nominal noise process are present.

Let us denote by $H(\mu\nu s_\ell^{-1})$ the entropy of the output process $[\mu\nu s_\ell^{-1}, D, Z]$, which is induced by the information carrying process $[\mu, A, X]$, the noise process $[\nu, B, Y]$, and the sliding block quantizer $s_\ell(\cdot)$.

As in [6,7,8], we will use the $\bar{\rho}$ distance as a measure of closeness between two stationary processes. As explained in [6], the $\bar{\rho}$ distance is the appropriate distance in this case, as opposed to the Prohorov distance. However, as in [6,7,8] our analysis will use the construction of sequences which imply Prohorov closeness between finite-dimension restrictions from the stationary measures. We emphasize this point here, to eliminate some objections as to the demanding properties of the $\bar{\rho}$ distance.

We now proceed with a definition for robust data smoothers.

Definition 4

Given a stationary and ergodic as well as smooth information carrying process $[\mu, A, X]$, given a high entropy stationary and ergodic noise process ν_o and a family M of stationary and ergodic processes that contains ν_o, the sequence $\{s_\ell\}$ of sliding block analog-to-digital converters is <u>robust at ν_o in M</u> iff:

Given $\epsilon > 0$, there is some $\delta > 0$ such that for all ℓ and all processes $\nu \epsilon M$:

$$\bar{\rho}(\nu_o, \nu) < \delta \rightarrow |H(\mu\nu_o s_\ell^{-1}) - H(\mu\nu s_\ell^{-1})| < \epsilon$$

We notice from definition 4 that our statistical contamination is included in the noise process only. That is, we focus on adjusting a sequence $\{s_\ell\}$ of parametric data smoothers designed for some fixed (but unknown) information carrying process and some well-known fixed noise process, in a way that entropy stability is guaranteed if the information carrying process remains the same but variations in the assumed noise process may exist. However, due to the lack of

precise statistical description on the information carrying process, our assumed statistical contamination can be at least partially carried over to this process. Our formalization is general enough to incorporate this generalization.

Now, as explained in [7,8], since the processes $[\mu \nu s_\ell^{-1}, D, Z]$ are finite alphabet processes (due to the fact that $s_\ell(\cdot)$ is an analog-to-digital converter), we can concentrate on $\{s_\ell\}$ sequences which guarantee $\bar{\rho}$ closeness of the processes $\mu \nu_o s_\ell^{-1}$ and $\mu \nu s_\ell^{-1}$. Such closeness will then imply the entropy closeness required by definition 4. To formalize this statement, we proceed with the following lemma:

Lemma 2

Given $[\mu, A, X]$, ν_o, M as in definition 4, and a sequence $\{s_\ell\}$ of sliding block analog-to-digital converters such that:

Given $\varepsilon > 0$, there is some $\delta > 0$ such that for all ℓ and all processes $\nu \varepsilon M$:

$$\bar{\rho}(\nu_o, \nu) < \delta \rightarrow \bar{d}(\mu \nu_o s_\ell^{-1}, \mu \nu s_\ell^{-1}) < \varepsilon$$

Then, the sequence $\{s_\ell\}$ is robust at ν_o in M.

In lemma 2, we notice that as in [7] we allow for possibly different distortion measures $d(\cdot, \cdot)$, $\rho(\cdot, \cdot)$ on the σ-algebras of the $\mu \nu s_\ell^{-1}$ and ν processes respectively. That is, $\bar{d}$ indicates rho bar distance defined through the distortion measure $d(\cdot, \cdot)$.

As in [7,8], if the data smoothers in the sequence $\{s_\ell\}$ are deterministic, each output Z_j in (6) assumes a unique value for every fixed value of the observed sequence $v_{j-\ell+m+1}^{j+m}$. If, on the other hand, each s_ℓ is stochastic, each observed sequence $v_{j-\ell+m+1}^{j+m}$ induces a measure.

b. Sufficient Conditions for Robust Data Smoothers

Our sufficient conditions will be basically the same as in [7]. We will present them here, referring to the proofs in [7], and then we will discuss some of the implications behind the distance measures we use.

Let the superposition relationship between the information carrying process $[\mu,A,X]$ and the noise process $[\nu,B,Y]$ be such that if $[\mu\nu,E,V]$ is the induced process, then,

$$\bar{\rho}(\nu_o,\nu) < \delta \rightarrow \bar{\rho}(\mu\nu_o,\mu\nu) < \delta \quad ; \quad \nu,\nu^o \varepsilon \, M. \tag{7}$$

Such superposition relationship is the additive.

Let E be the σ-algebra of sets from the alphabet E, and let γ be the metric on E. Let $\mathcal{D}$ be the σ-algebra of sets from the alphabet D, and let ξ be the metric on $\mathcal{D}$. Then, we will model each sliding block smoothing function s_ℓ, as a stationary, zero memory, (in general stochastic) channel $[E,s_\ell,\mathcal{D}]$ mapping sets from E^ℓ onto sets of $\mathcal{D}$, such that for some fixed integer m satisfying $m+1 \leq \ell$:

$$s^n_{\ell,v^{n-1+m}_{-\ell+m+1}}(z^{n-1}_o \varepsilon D_o \times D_1 \times \ldots \times D_{n-1}) = \prod_{j=o}^{n-1} s^1_{\ell,v^{j+m}_{j-\ell+m+1}}(Z_j \varepsilon D_j) \quad ; \quad \forall n, \, \forall D_j \varepsilon \, \mathcal{D}$$
$$; \quad i = o,\ldots,n-1 \tag{8}$$

Each s_{ℓ,v^ℓ} represents, in general, a measure for every given sequence v^ℓ. This measure is nontrivial if the mapping induced by s_ℓ is stochastic. s^1_{ℓ,v^ℓ} designates, in general, a conditional measure representing the transition probability of the channel in (8), conditioned on the sequence v^ℓ.

Let s^1_{ℓ,v^ℓ}, s^1_{ℓ,w^ℓ} be two conditional such measures conditioned on two different sequences v^ℓ, w^ℓ respectively.

Let $\Pi_{d,1} (s^1_{\ell,v^\ell}, s^1_{\ell,w^\ell})$ denote the Prohorov distance between the measures s^1_{ℓ,v^ℓ}, s^1_{ℓ,w^ℓ} ; where the distortion measure $d(\cdot,\cdot)$ is implied in the distance.

Let $\Pi_{d,k}$ indicate the k-dimensional Prohorov distance, where the distortion measure $d(\cdot,\cdot)$ is used.

Let B be the σ-algebra of sets from the alphabet B, and let the same metric γ as with the algebra E be the metric on B.

Let μ_{x^ℓ} be the empirical measure constructed from the sequence x^ℓ, as given by (2) in section 2, and let $\mu^k_{x^\ell}$ be its k-dimensional restrictions.

Then, as in [7], we can present the following definition:

<u>Definition 5</u>

i. The sequence $\{s_\ell\}$ of stochastic (in general) sliding block channels is <u>continuous</u>, if given $v^\ell \varepsilon E^\ell$, $\varepsilon > 0$, there exists $\delta = \delta(\ell, v^\ell, \varepsilon) > 0$ such that:

$$\gamma_\ell(v^\ell, w^\ell) < \delta \text{ implies } \Pi_{\varsigma,1} (s^1_{\ell,v^\ell}, s^1_{\ell,w^\ell}) < \varepsilon.$$

ii. The sequence $\{s_\ell\}$ of stochastic (in general) sliding block channels is <u>continuous at the measure v_0</u>, if given $\varepsilon > 0$, $\eta >$ there exist: integers k, ℓ_0, some $\delta > 0$, and for each $\ell > \ell_0$ some set $\Delta^\ell \varepsilon B^\ell$ with $v^\ell_0(\Delta^\ell) > 1-\eta$, such that: for each $x^\ell \varepsilon \Delta^\ell$, $y^\ell \varepsilon B^\ell$ with the property:

$$\Pi_{\gamma,k} (\mu^k_{x^\ell}, \mu^k_{y^\ell}) < \delta$$

it is implied that:

$$\Pi_{\xi,1} (s^1_{\ell,x^\ell}, s^1_{\ell,y^\ell}) < \varepsilon$$

The above definition is the same with definition 4 in reference [7]. Using the statements in this definition, we can now express the following theorem:

Theorem 2

Let μ be a fixed stationary and ergodic smooth information carrying process. Let ν_o be a given stationary and ergodic, high entropy, analog noise process. Let M be a given family of stationary and ergodic, high entropy, analog processes, that contains ν_o. Let a sequence $\{s_\ell\}$ of sliding block analog-to-digital converters be given. Let also the distortion measure $\rho(\cdot,\cdot)$ be used for the processes $\nu \in M$, $\mu\nu$, where $\mu\nu$ the process induced by the superposition of μ and ν. Let γ be the common metric on the σ-algebras of both ν and $\mu\nu$, and let the μ and ν superposition be such that:

$$\bar{\rho}(\nu_o,\nu) < \delta \to \bar{\rho}(\mu\nu_o,\mu\nu) < \delta \; ; \; \nu_o,\nu \in M \qquad (9)$$

Let the distortion measure $d(\cdot,\cdot)$ be used on the processes $\mu\nu s_\ell^{-1}$. Let $\rho(\cdot,\cdot)$, $d(\cdot,\cdot)$ be such that: Given $\epsilon > 0$, there exist $\delta_1 > 0$, $\delta_2 > 0$ such that:

$$\rho(X,Y) < \delta_1 \to \gamma(X,Y) < \epsilon$$

$$\xi(X,Y) < \delta_2 \to d(X,Y) < \epsilon$$

; where γ, ξ as in definition 5.

Then:

1. If $\{s_\ell\}$ is a sequence of stochastic sliding block channels that is continuous, then $\{s_\ell\}$ is also robust at ν_o in M (Def. 4) for finite length ℓ sliding block windows.

ii. If $\{s_\ell\}$ is a sequence of stochastic sliding block channels that is continuous at the measure ν_o, then $\{s_\ell\}$ is also robust at ν_o in M (Def. 4) for unbounded lengths ℓ of sliding block windows.

Due to the assumed property (9) in the theorem, the proof of it (including measurability of the sequence $\{s_\ell\}$), is exactly the same with the lengthy proof of theorem 2 in [7]. In fact, the statement in lemma 2 is proved there. This statement results in the property of robustness as expressed by definition 4.

Again, the sufficient conditions which guarantee the satisfaction of robustness, are such that they do not allow for deterministic sequences $\{s_\ell\}$ of sliding block analog-to-digital converting data smoothers. The conditions, imply again (as in [7,8]) that for the satisfaction of robustness, the sequence $\{s_\ell\}$ has to be a sequence of stochastic smoothers-converters.

There are a few points resulting from the presented analysis, which we feel that must be particularly emphasized. Thus, we include them in a separate section.

5.   Important Observations

1.   The analysis we have presented leads to the following general conclusion:
      <u>Data smoothers of high and robust performance, are a sequence of</u>
      <u>sliding block stationary stochastic analog-to-digital converters</u>.

      A sequence of sliding block data smoothers utilizes the available data sequences more efficiently. The analog-to-digital operation is needed for smoothing, or entropy compression. The stochastic nature of the smoothers guarantees performance stability in the presense of statistical contaminations.

      If the observation sequences, on which the smoothers operate, are of finite length (as it is usually the case), only continuity (Def. 5.i) of the smoothers, as real functions on the data, is needed for robustness.

2.   It has been felt by some researchers in the area of robustness that the $\bar{\rho}$ distance is too strong a measure as a contamination criterion on the space of the data processes. As emphasized in [6], the $\bar{\rho}$ distance is the only appropriate existing such measure for spaces of general stationary processes with memory.

The Prohorov distance does not provide then an appropriate upper bound on the distance between different finite restrictions of the processes. However, given two stationary processes μ_o, μ, if it is desired that their closeness be measured by the Prohorov distance $\Pi_k(\mu_o^k, \mu^k)$ between their k-dimensional restrictions μ_o^k and μ^k, the conditions in definition 5 and theorem 2, as well as the results in [6,7,8] hold as they are. The reason for that is, that in the proof of theorem 2 (found in [7]) only the Prohorov distance $\Pi_k(\mu_o^k, \mu^k)$ is used as a measure of closeness between the processes μ_o and μ.

3. There is an important side-result evolving from the presented theory, which we want to emphasize through a theorem. The result is asymptotic and it refers to stationary sliding block channels $[E, s_\ell, \mathcal{D}]$ with sliding block window of asymptotically large length.

Let us denote by $s(\mu\nu)$ the limit stationary and ergodic process induced by the process $\mu\nu$ and the sliding block stationary stochastic channel s_ℓ, for $\ell \to \infty$. Let us denote by $s^n(\mu\nu)$ n-dimensional restrictions of this process. Then, we can express the following theorem:

Theorem 3

Condition ii in definition 5 is sufficient and necessary for the satisfaction of the following property:

Given $\varepsilon > 0$, there exists $\delta > 0$ such that:

$$\bar{\rho}(\nu_o, \nu) < \delta \to \Pi_{d,1}(s^1(\mu\nu_o), s^1(\mu\nu)) < \varepsilon \; ; \; \forall \, \nu \varepsilon M.$$

Theorem 3 expresses a strong and important property. The property says that continuity at the measure ν_o (definition 4.ii) is also necessary (in addition to being sufficient for asymptotic entropy stability), if Prohorov closeness of the first-order restrictions of the processes induced by the sequence $\{s_\ell\}$ in the

limit ($\ell \to \infty$) is demanded. We call this property <u>consistency stability</u>.

The proof of theorem 3 can be found in the appendix. The same proof holds for the second part of theorem 2, in reference [8].

6. The Reconstruction of Continuous – Time Waveforms

As stated by Huber [2], one of the important, possibly subsequent operations of the data smoothers, is the reconstruction of continuous-time waveforms $f(t)$ from a discrete-time sequence x generated by the information carrying process $[\mu, A, X]$.

In this section, we will show that such reconstruction is asymptotically possible only if the analog-to-digital converters s_ℓ are stochastic. This result presents an additional argument in favor of stochastic analog-to-digital sliding-block data smoothers, which also satisfy the conditions of definition 5.

Let x be a discrete-time sequence generated by the information carrying process $[\mu, A, X]$. Let $[\nu_x, C, W]$ be the process induced by the sequence x and the noise process $[\nu, B, Y]$. Let $\{s_\ell\}$ be a sequence of sliding-block analog converters operating on the sequence w.

Let $f(t)$; $-\infty < t < \infty$ be the continuous-time waveform whose sampled version is the sequence x. Let these samples be in distance u from each other, so that the kth element x_k from the sequence x is given by $x_k = f(\frac{k}{u})$. Let $g(x_k)$ be the quantized version of the element x_k, as induced by the sequence $\{s_\ell\}$.

Let $\{u_n\}$ be a sequence monotonically decreasing to zero (for n increasing), and let $\{h_k(u_n, t)\}$ be a sequence of kernel functions satisfying the properties in [9], where the sequence is generated by varying k.

Define

$$f_{u_n}(t) = \sum_{k=-\infty}^{\infty} x_k \, h_k(u_n, t) \tag{9}$$

Then, it is well-known that $f_{u_n}(t) \to f(t)$.
$$ n \to \infty$$

Define

$$g_{u_n}(t) = \sum_{k=-\infty}^{\infty} g(x_k)\, h_k(u_n, t) \tag{10}$$

Then, we can express the following lemma:

Lemma 3

If $f(t)$ is a bounded, uniformly continuous function on $(-\infty, \infty)$ and $g(x_k)$ is a deterministic analog-to-digital mapping, then the function $g_{u_n}(t)$ <u>does not</u> converge to $f(t)$ for $n \to \infty$.

Proof

Suppose that $g_{u_n}(t)$ converged to $f(t)$ for some t. Then, given $\varepsilon > 0$, there exists n_o such that:

$$\left| f_{u_n}(t) - g_{u_n}(t) \right| < \varepsilon \; ; \; \forall\, n > n_o. \tag{11}$$

Let

$$y = \{\ldots,\, y_{-1},\, y_o,\, y_1,\, \ldots\}$$

; where $y_k = g(x_k)$.

Let μ_x, μ_y be the empirical measures from the sequences x and y respectively, as in (2). Then, due to theorem 3, property (11) can be satisfied if and only if some $\delta > 0$ can be found such that:

$$\Pi_{\gamma, 1}\, (\mu_x^1,\, \mu_y^1) < \delta \text{ implies } \left| f_{u_n}(t) - g_{u_n}(t) \right| < \varepsilon \; ; \; \forall\, n > n_o \tag{12}$$

for all x sequences within some compact space of A^∞ which covers most of A^∞. Such a compact space clearly exists for bounded $f(t)$ functions.

But if $g(x_k)$ is deterministic, it is also discontinuous. Therefore, property (12) can not be satisfied.

Let now the mapping $g(x_k)$ be stochastic. Specifically, let $g(x_k)$ be any measure continuous in x_k. That is, given $\epsilon > 0$ and x_k, there exists some $\delta = \delta(\epsilon, x_k) > 0$ such that

$$|x_k - z| < \delta \rightarrow \Pi_{\xi,1} (g(x_k), g(z)) < \epsilon.$$

Such a $g(\cdot)$ generates a sequence of measures $g_{u_n}(t)$, for every fixed t and varying n.

Define

$$G_{u_n}(t) = \sum_{k=-\infty}^{\infty} E\{g(x_k)\} \cdot h_k(u_n,t) \tag{13}$$

where $E\{g(x_k)\}$ the expected value of $g(x_k)$.

Define

$$G(t) = \lim_{n \to \infty} G_{u_n}(t) \tag{14}$$

and let $G(t)$ exist.

Then, $G(t)$ is uniquely determined through the choise of the $g(x_k)$ measures.

Let the measures $g(x_k)$ be such that the function $G(t)$ in (14) is a one-to-one, continuous, and monotone function of $f(t)$. That is, $f(t) = G^{-1}(f(t))$, and given $\epsilon > 0$, there exists some $\delta > 0$, such that:

$$|G(t_1) - G(t_2)| < \delta \rightarrow |f(t_1) - f(t_2)| < \epsilon$$

Then, we can express the following theorem.

Theorem 4

Let the analog-to-digital mapping $g(x_k)$ be stochastic, continuous as a function of x_k, and such that it induces a function $G(t)$ which is a one-to-one, continuous, and monotone function of $f(t)$. Then, the function

$$g_{u_n}(t) = \sum_{k=-\infty}^{\infty} g(x_k)\, h_k(u_n, t)$$

converges for $n \to \infty$ to $G(t)$ almost everywhere for all t. Therefore, $G^{-1}(f(t))$ converges then to $f(t)$ a.e. for all t.

Proof

Let n_1, n_2 be given. Define

$$x_k = f\left(\frac{k}{u_{n_1}}\right) \,, \quad y_k = f\left(\frac{k}{u_{n_2}}\right)$$

$$x = \{\ldots, x_{-1}, x_o, x_1, \ldots\}$$

$$y = \{\ldots, y_{-1}, y_o, y_1, \ldots\}$$

Due to the nature of the kernel functions [as in 9], and the boundness and uniform continuity of $f(t)$, we have:

There exists some n_o and some compact space Δ in A^∞ which covers most of A^∞, such that condition ii in definition 5 is satisfied for all n_1, $n_2 > n_o$ and $x \varepsilon \Delta$, $y \varepsilon A^\infty$. Specifically, given n_1, $n_2 > n_o$, $x \varepsilon \Delta$ and $\varepsilon > 0$, there exists some $\delta > 0$ independent of x, such that:

$$\Pi_{\gamma,1}\,(\mu_x^1, \mu_y^1) < \delta$$

implies $\Pi_{\xi,1}(g_{u_{n_1}}(t), g_{u_{n_2}}(t)) < \varepsilon$; $\forall\, t$; $\forall\, n_1, n_2 > n_o$

But since $g_{u_{n_1}}(t)$, $g_{u_{n_2}}(t)$ are bounded, we also have then:

$$V_\xi(g_{u_{n_1}}(t), g_{u_{n_2}}(t)) < \varepsilon \; ; \; \forall t \; ; \; \forall\, n_1, n_2 > n_o$$

; where $V_\xi(\cdot,\cdot)$ denotes the Vasershtein distance defined through the metric ξ.

But Vasershtein closeness implies convergence to the mean of $G(t)$ a.e. for all t.

The proof is now complete.

Lemma 3 and theorem 4 prove that the design of stochastic analog-to-digital smoothing functions is necessary and sufficient for the asymptotic reconstruction of the f(t) waveform from its discrete-time version x.

Appendix

Proof of theorem 3

First we will prove the sufficient of the condition in definition 5.ii.
Then, we will prove the necessary.

1. Let the condition in definition 5.ii. be true.

   In this case, we only have to prove the existence of $s^1(\mu\nu)$ for each $\nu\epsilon M$; that is consistency for some $s^1(\mu\nu)$.

   Then, since for all $\ell > \ell_0$, given $\epsilon > 0$, there is $\delta > 0$ such that:

   $$\bar{\rho}(\nu_0,\nu) < \delta \to \bar{d}(\mu\nu_0,\mu\nu) < \epsilon \; ; \; \nu_0,\nu\epsilon M$$

   if the condition in definition 5.ii, holds, then clearly:

   $$\bar{\rho}(\nu_0,\nu) < \delta \to \Pi_{d,1}(s^1(\mu\nu_0), \; s^1(\mu\nu)) < \epsilon$$

   also, since the $\bar{d}$ distance bounds from above the Prohorov distance $\Pi_{d,k}$, for all k.

   To prove consistency at some $\nu\epsilon M$, we only have to prove that for any $\delta > 0$ and for some $s^1(\mu\nu)$, we have:

   $$\lim_{\ell\to\infty} \mu\nu(x: \Pi_{d,1}(s^1_{\ell,x}\ell, \; s^1(\mu\nu)) > \delta) = 0$$

   But this results exactly the same way and by the same construction
   of sets as in the proof of lemma 5.2. in [6], in conjuction with the
   fact that the condition in definition 5.ii. implies:

   $$\left. \begin{array}{l} \Pi_{\gamma,k} (\mu^k_{x}\ell, \; \mu\nu^k) < \delta \\[2ex] \Pi_{\gamma,k} (\mu^k_{y}m, \; \mu\nu^k) < \delta \end{array} \right\} \to \Pi_{d,1}(s^1_{\ell,x}\ell, \; s^1_{m,y}m) < \epsilon.$$

2. Assume that the following property is true:

Given $\epsilon > 0$, there exists $\delta > 0$, such that:

$$\bar{\rho}(\nu_o, \nu) < \delta \to \Pi_{d,1}(s^1(\mu\nu_o), s^1(\mu\nu)) < \epsilon \; ; \; \forall \; \nu_o, \nu \epsilon M \qquad (A.1)$$

The implicit assumption behind the above property is existence of consistency. In particular, it is implied that the sequence $\{s_\ell\}$ is consistent at all measures ν in a neighborhood of ν_o, within M.

But $\bar{\rho}(\nu_o, \nu) < \delta$ implies

$$\Pi_{\rho,k}(\nu_o^k, \nu^k) < \delta \; ; \; \forall \; k$$

Now, suppose that the condition in definition 5.ii. does not hold. Then, there is no k and no ℓ_o such that

$$\Pi_{\rho,k}(\nu_o^k, \nu^k) < \delta \to \Pi_{d,1}(\mu\nu_o^1 s_\ell^{-1}, \mu\nu^1 s_\ell^{-1}) < \epsilon \; ; \; \forall \; \ell > \ell_o$$

so (A.1) can not be true.

Therefore the nonholding of the condition in definition 5.ii. presents a contradiction. Thus, necessity holds.

References:

[1] J. W. Tukey (1977), Exploratory Data Analysis, Addison-Wesley, Reading, Mass.

[2] P. Huber (1979), "Robust Smoothing" in Robustness and Statistics, edited by
     R. L. Launer and G. Wilkinson, Academic Press.

[3] C. L. Mallows (1979), "Some Theory of Nonlinear Smoothers," to appear in
     the Annals of Statistics.

[4] C. L. Mallows (1979), "Resistant Smoothing" to appear in Proc. Heidelberg
     Workshop on Smoothing Tech. for Curve Est. (T. Gasser and M. Rosenblatt,
     editors).

[5] R. D. Martin (1979), "Approximate Conditional-Mean Type Smoothers and Inter-
     polators," Invited talk, to appear in Proc. Heidelberg Workshop on Smoothing
     Tech. for Curve Est. (T. Gasser and M. Rosenblatt, editors).

[6] P. Papantoni-Kazakos and R. M. Gray (1979), "Robustness of Estimation on
     Stationary Observations", Annals of Probability, Vol. 7, No. 6, December,
     pp. 989-1002.

[7] P. Papantoni-Kazakos (1979), "Sliding Block Encoders that are Rho-Bar
     Continuous Functions of their Input," Tech. Rep. TR-79-11, Dept. of
     Electrical Engineering and Computer Science, The University of Connecticut,
     Storrs, CT., August, and to appear in the IEEE Trans. Inf. Th., May 1981.

[8] P. Papantoni-Kazakos (1980), "Stochastic Quantization for Performance
     Stability," Tech. Rep. TR-80-6, Dept. of Electrical Engineering and Computer
     Science, The University of Connecticut, Storrs, CT., March, and to be
     presented at the 1981 International Symposium on Information Theory.

[9] E. Masry and S. Cambanis, "Consistent Estimation of Continuous-Time Signals
     from Quantized Noisy Samples," 1979 Annual Conference on Information Sciences
     and Systems, Johns Hopkins Univ.